OPIC
Office de la propriété intellectuelle du Canada

CIPO
Canadian Intellectual Property Office

Ottawa Hull K1A 0C9

(11) (C) **1,340,280**

(21) 492,345

(22) 1985/10/04

(45) 1998/12/22

(52) 195-1.23
C.L. CR. 195-1.22
195-1.29
195-1.35

(51) Int.Cl.[6] C12N 15/67; C12N 15/11

(19)(CA) **CANADIAN PATENT** (12)

# (54) Synthetic Signal Sequence for the Transport of Proteins in Expression Systems

(72) Engels, Joachim , Germany (Federal Republic of)
Leineweber, Michael , Germany (Federal Republic of)
Uhlmann, Eugen , Germany (Federal Republic of)
Wetekam, Waldemar , Germany (Federal Republic of)

(73) HOECHST AKTIENGESELLSCHAFT , Germany (Federal Republic of)

(30) (DE) Germany (Federal Republic of) P 34 36 818.3 1984/10/06

(57) 16 Claims

1340280

HOE 84/F 238

Abstract of the disclosure:

Synthetic signal sequence for the transport of proteins in expression systems

The DNA of a natural signal sequence is modified by incorporation of cleavage sites for endonucleases and can thus be incorporated in any desired vectors by the modular construction principle. The vectors modified in this way then bring about transport of the coded protein out of the cytoplasm.

# A synthetic signal sequence for the transport of proteins in expression systems

In the cell, proteins are synthesized on the ribosomes which are located in the cytoplasm. Proteins which are transported out of the cytoplasm carry on the amino terminal end a relatively short peptide chain which is eliminated enzymatically on passage through the cytoplasmic membrane, whereupon the mature protein is produced. This short peptide sequence is called a "signal peptide" or a presequence or leader sequence.

The signal sequence located at the amino terminal end has already been characterized for a large number of secretory proteins. In general, it is composed of a hydrophobic region of about 10 to 20 amino acids, which is called the core and to whose amino terminal end a short peptide sequence (the pre-core) is bonded, this usually having one positively charged amino acid (or several). Between the carboxy terminal end of the hydrophobic region and the amino terminal end of the mature transported protein there is a short peptide sequence (the post-core) which contains the splice site and ensures that the spatial arrangement is favorable.

It is known, from U.S. Patent 4,411,994, to couple the gene for a protein which is to be expressed with a bacterial gene which codes for an extracellular or periplasmic carrier protein in order thus to bring about the transport of the desired protein out of the cytoplasm. It is necessary for this process to isolate a bacterial gene, which is intrinsic to the host, for a periplasmic, outer membrane protein or an extracellular protein. This gene is then cut with a restriction enzyme, the gene for the protein which is to be transported is inserted into the cut which has been produced, and the host cell is transformed

with a vector which contains the fusion gene thus formed. The isolation of the natural gene and its characterization for the selection of suitable cleavage sites is extremely complex. This complexity is avoided according to the invention by making use of a synthetic signal sequence.

Thus the invention relates to a synthetic signal sequence for the transport of proteins in expression systems, which comprises DNA essentially corresponding to a natural signal sequence but having one or more cleavage sites for endonucleases which are not present in the natural DNA.

The invention further relates to DNA of the Formula I (see page 17).

The invention further relates to a process for the transport expression of eukaryotic, prokaryotic or viral proteins in prokaryotic and eukaryotic cells, which comprises coupling the gene for the protein which is to be transported onto a DNA sequence as described above, incorporating this fusion gene into a vector, and transforming therewith a host cell which transports the expressed protein out of the cytoplasm.

The invention further relates to a hybrid vector comprising a DNA sequence as described above and a host organism containing such vector.

The invention will now be described in further detail by reference to the appended drawings:

Figure 1 shows the digestion of the plasmid pBR 322 with the restriction endonucleases EcoR I and Pvu II and then the filling in of the EcoR I cleavage site.

Figure 2 shows the plasmid pUC 9 containing the monkey preproinsulin DNA and the reaction sequence for the construction of the proinsulin DNA fragment.

Figure 3 shows the ligation of the chemically synthesized regulation region with the proinsulin DNA fragment.

Figure 4 shows how the hybrid plasmid pVI 6 is obtained.

Figure 5 shows the plasmid pWI P1 having a DNA sequence I integrated in the correct direction of reading to the proinsulin gene.

The DNA should "essentially" correspond to that of a natural signal sequence. This is to be understood to mean that the expressed signal peptide is substantially or completely identical to the natural signal peptide, in the latter case therefore the only difference existing at the DNA level is that the synthetic DNA has at least one cleavage site that the natural DNA sequence does not contain. This incorporation of the cleavage site according to the invention thus means that there is a, more or less extensive, difference from the natural sequence, it being necessary under certain circumstances to have recourse to codons which are known to be less preferred by the particular host organism. However, surprisingly, this is not associated with any expression disadvantage. On the contrary, the specific "making to measure" of the synthetic gene is associated with so many advantages that any disadvantage owing to the use of "unnatural" codons is, in general, overcompensated by far. In fact, it has emerged that replacement of the start codon GTG, which occurs in the gene for alkaline phosphatase in E. coli, by ATG leads to a great increase in expression. A particular advantage of the invention is that the host cell has to produce less ballast protein because the gene which is to be expressed can be directly linked to the 3' end of

the synthetic DNA signal sequence. Furthermore, advantages accrue in so far as it is possible in the construction of the synthetic DNA to provide DNA sequences, which protrude at the ends, for certain restriction recognition sites which allow cloning of this sequence and, in the case of disparate recognition sites, permit defined incorporation into a cloning vector. This makes possible incorporation to any desired vectors by the "modular construction principle".

Internal recognition sites for restriction enzymes permit any desired homologous or heterologous genes to be coupled on in the correct reading frame. It is also possible via these internal cleavage sites to introduce in a straightforward manner modifications in the DNA of the signal sequences, which lead to presequences which do not occur in nature.

These internal cleavage sites are advantageously placed in the regions upstream and downstream of the hydrophobic region, in particular in the post-core region, it being possible to modify the splice site and/or its adjacent region. Of course, it is also possible to modify the core region in a manner known per se.

Taking known rules into account (G. von Heijne, J. Mol. Biol. 173 (1984) 243-251) it is possible, via suitable cleavage sites in the gene section which codes for the carboxy terminal part of the prepeptide, to plan the signal peptidase splice site in such a manner that there is expression not of a fusion protein but directly of the desired, generally eukaryotic, peptide in its natural form. In general, genes of natural origin do not allow processing of this type.

Suitable signal sequences are in principle all signal sequences known from the literature (M.E.E. Watson; Nucleic Acids Res. 12 (1984), 5145 - 5164), modifications thereof and "idealized" signal sequences derived therefrom

(D. Perlman and H.O. Halvorson; J. Mol. Biol. 167 (1983), 391 - 409).

Preferred host organisms are E. coli, Streptomyces, Staphylococcus species, such as S. aureus, Bacillus species, such as B. subtilis, B. amyloliquifaciens, B. cereus or B. licheniformis, Pseudomonas, Saccharomyces, Spodoptera frugiperda and cell lines of higher organisms, such as plant or animal cells.

In principle, it is possible to obtain by transport expression all those proteins of prokaryotic or eukaryotic origin which can pass through the membrane. However, peptide products which are of pharmaceutical significance, such as hormones, lymphokines, interferons, blood-coagulation factors and vaccines, which in nature are also coded as peptides with an amino-terminal presequence are preferred. However, in the prokaryotic host organisms this eukaryotic presequence is not, as a rule, eliminated by the signal peptidases intrinsic to the host.

In E. coli, the genes for the periplasmic and outer-membrane proteins are suitable for transport expression, the former directing the product into the periplasm whereas the latter tend to direct onto the outer membrane.

The example which is given is the DNA signal sequence of the periplasmic protein alkaline phosphatase, which is very readily expressed in E. coli, but there is no intention to restrict the invention to this.

The presequence including the first twenty amino acids of alkaline phosphatase of E. coli is shown below:

```
 1                   5                  10
Met-Lys-Gln-Ser-Thr-Ile-Ala-Leu-Ala-Leu-Leu-Pro-Leu-Leu-

15                  20                  25
Phe-Thr-Pro-Val-Thr-Lys-Ala-Arg-Thr-Pro-Glu-Met-Pro-Val-
                           ↑

30                  35                  40
Leu-Glu-Asn-Arg-Ala-Ala-Gln-Gly-Asn-Ile-Thr-Ala-Pro
```

↑ = preferred splice site of the signal peptidase

It has emerged that up to about 40, usually about 20, additional amino acids of the mature protein suffice for correct processing. However, in many cases fewer additional amino acids also suffice, for example about 10, advantageously about 5. Since a shorter protein chain means less stress on the protein biosynthesis system of the host cell, an advantageous embodiment of the invention is set out in DNA sequence I (see page 17) which codes for the presequence of alkaline phosphatase and an additional 5 amino acids of the perfect protein. Apart from a few triplet modifications - namely those which introduce unique restriction enzyme cleavage sites and replace the start codon GTG by ATG - DNA sequence I corresponds to the natural sequence for alkaline phosphatase. At the ends of the coding strand are located protruding DNA sequences corresponding to the restriction endonuclease EcoR I, which permit incorporation into conventional cloning vectors, for example the commercially available plasmids such as pBR 322, pUC 8 or pUC 12. In addition, a number of other unique cleavage sites for restriction enzymes have been incorporated within the gene of DNA sequence I, and these, on the one hand, make it possible to couple heterologous genes onto the correct site and in the desired reading frame and, on the other hand, permit modifications to be carried out:

| Restriction enzyme | Cut after nucleotide No. (in the coding strand) | |
|---|---|---|
| Sau 3 A | 19 | |
| Pvu I | 22 | |
| Hpa II | 54 | ) (present in the |
| Nci I | 54 | ) natural gene) |
| Alu I | 66 | |
| Hph I | 68 | |
| Ava II | 70 | |

Of course, it is also possible to construct the protruding

sequences in such a manner that they correspond to different restriction enzymes, and this then permits incorporation into suitable vectors in a defined orientation. In this context, the expert will give consideration to whether the complexity associated with the construction of the gene and its specific incorporation is more important than the additional work of selection associated with incorporation in both orientations when the protruding ends are identical.

DNA sequence I can be constructed of 6 oligonucleotides 26 - 31 bases in length by first synthesizing them chemically and then linking them enzymatically via sticky ends of 6 nucleotides. Incorporation of the synthetic gene into cloning vectors, for example into the commercially available plasmids mentioned, is carried out in a manner known per se.

As an example for the expression of a eukaryotic gene in E. coli using a presequence according to the invention, the synthesis of monkey proinsulin is described below: a DNA sequence is constructed in which the DNA sequence I, followed by the proinsulin gene (W. Wetekam et al., Gene 19 (1982) 179-183), is located on a connecting recognition site for EcoR I and downstream of a chemically synthesized regulation region, composed of a bacterial promoter, a lac operator and a ribosomal binding site (German Patent Application P 34 30 683.8), and 6 - 14 nucleotides away from the ribosomal binding site. The expressed proinsulin fusion peptide contains an additional 9 amino acids on the amino terminal end, and these can be eliminated enzymatically or chemically.

The incorporation of the synthetic gene into pUC 8 and the construction of expression plasmids which contain the eukaryotic genes coupled to DNA sequence I are carried out in a manner known per se. In this context, reference may be made to the textbook by Maniatis (Molecular Cloning, Maniatis et al., Cold Spring Harbor, 1982). The

transformation of the hybrid plasmids thus obtained into suitable host organisms, advantageously E. coli, is likewise known per se and is described in detail in the abovementioned textbook. The isolation of the expressed proteins and their purification is likewise described.

In the examples which follow some more embodiments of the invention are specifically illustrated, from which is evident to the expert the large number of possible modifications (and combinations). Unless otherwise specified, percentage data in these examples relate to weight.

Examples

1. Chemical synthesis of a single-stranded oligonucleotide

The synthesis of the structural units of the gene is illustrated by the example of structural unit Ia of the gene, which comprises nucleotides 1 - 29 of the coding strand. The nucleoside at the 3' end, in the present case therefore guanosine (nucleotide No. 29), is covalently bonded via the 3'-hydroxy group, by known methods (M.J. Gait et al., Nucleic Acids Res. 8 (1980) 1081 - 1096) to silica gel (FRACTOSIL*, supplied by Merck). For this purpose, first the silica gel is reacted with 3-triethoxysilylpropylamine with elimination of ethanol and formation of a Si-O-Si bond. The guanosine is reacted as the $N^2$'-isobutyryl-3'-O-succinoyl-5'-dimethoxytrityl ether with the modified carrier in the presence of paranitrophenol and N,N'-dicyclohexylcarbodiimide, the free carboxy group of the succinoyl group acylating the amino radical of the propylamine group.

In the synthetic steps which follow, the base component is used as the monomethyl ester of the 5'-O-dimethoxytritylnucleoside-3'-phosphorous acid dialkylamide or chloride, the adenine being in the form of the $N^6$-benzoyl compound, the cytosine being in the form of the $N^4$-benzoyl compound, the guanine being in the form of the $N^2$-iso-

* Trademark

butyryl compound, and the thymine, which contains no amino group, being without a protective group.

50 mg of the polymeric carrier containing 2 µmol of bound guanosine are treated successively with the following agents:

a) nitromethane
b) saturated zinc bromide solution in nitromethane containing 1% water
c) methanol
d) tetrahydrofuran
e) acetonitrile
f) 40 µmol of the appropriate nucleoside phosphite and 200 µmol of tetrazole in 0.5 ml of anhydrous acetonitrile (5 minutes)
g) 20% acetic anhydride in tetrahydrofuran containing 40% lutidine and 10% dimethylaminopyridine (2 minutes)
h) tetrahydrofuran
i) tetrahydrofuran containing 20% water and 40% lutidine
j) 3% iodine in collidine/water/tetrahydrofuran in the ratio by volume 5 : 4 : 1
k) tetrahydrofuran and
l) methanol.

In this context, the term "phosphite" is to be understood to be the monomethyl ester of the deoxyribose-3'-monophosphorous acid, the third valency being saturated by chloride or a tertiary amino group, for example a morpholino radical. The yields in each synthetic step can be determined after the detritylation reaction (b) in each case by spectrophotometry, measuring the absorption of the dimethoxytrityl cation at a wavelength of 496 nm.

When the synthesis of the oligonucleotide is complete, the methyl phosphate protective groups on the oligomer are eliminated using p-thiocresol and triethylamine. The oligonucleotide is then removed from the solid carrier by treatment with ammonia for 3 hours. Treatment of the

oligomers with concentrated ammonia for 2 to 3 days quantitatively eliminates the amino protective groups on the bases. The crude product thus obtained is purified by high-pressure liquid chromatography (HPLC) or by polyacrylamide gel electrophoresis.

The other structural units Ib - If of the gene are synthesized entirely correspondingly, their nucleotide sequences being evident from DNA sequence II (see page 18).

2. Enzymatic linkage of the single-stranded oligonucleotides to give DNA sequence I

The terminal oligonucleotides Ia and If are not phosphorylated. This prevents oligomerization via the protruding ends. For the phosphorylation of oligonucleotides Ib, Ic, Id and Ie, in each case 1 nmol of these compounds is treated with 5 nmol of adenosine triphosphate and 4 units of T4 polynucleotide kinase in 20 µl of 50 mM tris.HCl buffer (pH 7.6), 10 mM magnesium chloride and 10 mM dithiothreitol (DTT) at 37°C for 30 minutes. The enzyme is inactivated by heating at 95°C for 5 minutes. The oligonucleotides Ia to If are then combined and hybridized to give the double strand by heating them in a 20 mM KCl solution and then slowly (over the course of 2 hours) cooling to 16°C. The ligation to give the DNA fragment according to DNA sequence I is carried out by reaction in 40 µl of 50 mM tris.HCl buffer (20 mM magnesium chloride and 10 mM DTT) using 100 units of T4 DNA ligase, at 15°C over the course of 18 hours.

The purification of the gene fragment is carried out by gel electrophoresis on a 10% polyacrylamide gel (without addition of urea, 20 x 40 cm, 1 mm thick), the marker substance used being ØX 174 DNA (supplied by BRL) cut with Hinf I, or pBR 322 cut with Hae III.

3. Incorporation of the gene fragment in pUC 8

The commercially available plasmid pUC 8 is opened in a known manner and in accordance with the manufacturer's data using the restriction endonuclease EcoR I. The digestion mixture is fractionated by electrophoresis on a 5% polyacrylamide gel in a known manner, and the DNA is visualized by staining with ethidium bromide or by radioactive labeling ("Nick translation" method of Maniatis, loc. cit.). The plasmid band is then cut out of the acrylamide gel and separated from the polyacrylamide by electrophoresis.

4. Incorporation of DNA sequence I into an expression plasmid

The expression plasmid pWI 6 having the information for monkey proinsulin is constructed as follows:

10 µg of the plasmid pBR 322 are digested with the restriction endonucleases EcoR I and Pvu II and then the EcoRI cleavage site is filled in a fill-in reaction using Klenow polymerase. Following fractionation by gel electrophoresis in a 5% polyacrylamide gel, the plasmid fragment of length 2293 Bp can be obtained by electroelution (Figure 1).

The monkey preproinsulin DNA integrated in the plasmid pBR 322 (Wetekam et al., Gene 19 (1982) 179 - 183) is isolated by digestion using the restriction endonucleases Hind III and Mst I (as a fragment of about 1250 Bp) and recloned into the plasmid pUC 9 as follows: the plasmid pUC 9 is cleaved with the enzyme Bam HI, the cleavage site is filled in a standard fill-in reaction using Klenow polymerase ("large fragment"), subsequent cleavage with the restriction enzyme Hind III is carried out, and the DNA is separated from the other DNA fragments by gel electrophoresis in a 5% polyacrylamide gel. The isolated insulin DNA fragment of length about 1250 Bp is integrated

into the opened plasmid.

To remove the untranslated region and the presequence, the pUC 9 plasmid thus modified is digested with Hae III, and the fragment of length 143 Bp is digested with Bal 31 under limiting enzyme conditions to eliminate the last two nucleotides from the presequence. This results in the first codon on the amino terminal end being TTT, which represents phenylalanine as the first amino acid of the B chain.

An adaptor which is specific for Eco RI is now ligated onto this fragment in a blunt-end ligation reaction:

a) 5' AAT TAT GAA TTC GCA ATG
Eco RI TA CTT AAG CGT TAC

b) 5' AAT TAT GAA TTC GCA AGA
Eco RI TA CTT AAG CGT TCT

In order to prevent polymerization of the adaptors they are used unphosphorylated in the ligation reaction (this being indicated in the figures by Eco RI$^-$, in the same way as recognition sequences inactivated by, for example, filling in). The adaptor a) has a codon for methionine at the end, and the adaptor b) has a codon for arginine. Thus, the gene product obtained by variant a) is amenable to removal of the bacterial contribution by cleavage with cyanogen bromide, whereas variant b) allows trypsin cleavage.

The ligation product is digested with Mbo II. After fractionation by gel electrophoresis, a DNA fragment of length 79 Bp having the information for amino acids Nos. 1 to 21 of the B chain is obtained.

The gene for the remaining information for the proinsulin molecule (including a G-C sequence from the cloning and 21 Bp from the pBR 322 connected to the stop codon) is

obtained from the pUC 9 plasmid having the complete information for monkey preproinsulin by digestion with Mbo II/Sma I and isolation of a DNA fragment of length about 240 Bp. The correct ligation product of length about 320 Bp (including the adaptor of 18 Bp) is obtained by ligation of the two proinsulin fragments. This proinsulin DNA fragment thus constructed can now be ligated together with a regulation region via the Eco RI negative cleavage site.

Figure 2 shows the entire reaction sequence, where A, B and C denote the DNA for the particular peptide chains of the proinsulin molecule, Ad denotes the (dephosphorylated) adaptor (a or b) and Pre denotes the DNA for the presequence of monkey preproinsulin.

A chemically synthesized regulation region composed of a recognition sequence for Bam HI, the lac operator (O), a bacterial promoter (P) and a ribosomal binding site (RB), and having an ATG start codon, 6 to 14 nucleotides away from the RB and having a connected recognition sequence for Eco RI (Figure 3) is ligated, via the common Eco RI overlapping region, with the proinsulin gene fragment obtained according to the previous example. It is advantageous to choose the following synthetic regulation region (DNA sequence IIa from Table 2, corresponding to German Patent Application P 34 30 683.8):

```
5' GATCCTAAATAAATTCTTGACATTTTTTAAA 3'
3'     GATTTATTTAAGAACTGTAAAAAATTT 5'

   (Bam HI)                    P

5' TAATTTGGTATAATGTGTGGAATTGTGAGCG 3'
3' ATTAAACCATATTACACACCTTAACACTCGC 5'

                        O

5' GAATAACAATTTCACAGAGGATCTAG      3'
3' CTTATTGTTAAAGTGTCTCCTAGATCTTAA 5'

                      RB     (Eco RI)
```

The other synthetic regulation regions specified in Table 2 can be used likewise. However, it is also possible to choose a natural or derived (Perlman et al., loc. cit.) signal sequence known from the literature.

TABLE 2

Synthetic regulation region (coding strand):

5' GGATCCTAAATAAATTCTTGACATTT1TTAA2TAATTTGGTATAATGT3T

4GAATTG5GAGCG6T7ACAATT8C9C10G11G12T13TA14TT15 (ATG) 3'

1 = T or G
2 = A or C
3 = G or C
4 = G or A
5 = T or C
6 = C, GA or GAA

7 = A or C
8 = T or direct bond
9 = A or TAGA
10 = A, TTTAAA, AAGCTT or AAGCTA

11 = AG or GA
12 = A or G
13 = C or T
14 = GAA or AGC
15 = C or direct bond

DNA sequences IIa-h:

| | 1 | 2 | 5 | 6 | 7 | 8 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| IIa | T | A | T | GAA | A | T | A | AG | A | C | GAA | C |
| b | T | A | T | GAA | A | T | TTTAAA | AG | A | C | GAA | C |
| c | G | C | T | GAA | A | T | TTTAAA | AG | A | C | GAA | C |
| d | G | C | T | GAA | A | - | AAGCTT | AG | A | C | GAA | C |
| e | G | C | C | GAA | A | - | AAGCTT | AG | A | C | GAA | C |
| f | G | C | T | C | C | - | AAGCTT | AG | A | C | GAA | C |
| g | G | C | T | GA | C | - | AAGCTT | AG | A | C | GAA | C |
| h | G | C | T | GA | C | - | AAGCTA | GA | G | T | AGC | - |

3 = G
4 = G
9 = A

Following double digestion with Sma I/Bam HI and a fill-in reaction of the Bam HI cleavage site with the Klenow fragment, the ligation product (about 420 Bp) is isolated by gel electrophoresis.

The fragment thus obtained can then, by a blunt-end ligation, be ligated into the pBR 322 part-plasmid of Figure 1 (Figure 4). The hybrid plasmid pWI 6 is obtained.

After transformation into the E. coli strain HB 101 and selection on ampicillin plates, the plasmid DNA of individual clones was tested for the integration of a 420 Bp fragment having the regulation region and the proinsulin gene shortened by Bal 31. In order to demonstrate the correct shortening of the proinsulin gene by Bal 31 (Figure 2), the plasmids having the integrated proinsulin gene fragment were sequenced starting from the Eco RI cleavage site. Of 60 sequenced clones, three had the desired shortening by two nucleotides (Figure 4).

1 µg of the plasmid pWI 6 is cut with the restriction enzyme Eco RI and then ligated together in the presence of 30 ng of DNA sequence I, at 16°C in 6 hours. After transformation into E. coli HB 101, plasmids are isolated from individual clones and tested for integration of DNA sequence I by means of restriction enzyme analysis. 7% of the clones contained the plasmid pWI 6 with integrated DNA sequence I.

The direction of this integration reaction can be unambiguously determined by standard methods of restriction enzyme analysis via double digestion with Hind III/ Pvu I. The plasmid pWI 6 having a DNA sequence I integrated in the correct direction of reading to the proinsulin gene is shown as pWIP 1 in Figure 5.

This plasmid can then be transformed into various E. coli strains in order to test the synthetic capacity of the individual strains.

The expression of the presequence-proinsulin gene fusion in E. coli is determined as follows:

1 ml of a bacterial culture induced with IPTG (isopropyl β-D-thiogalactopyranoside) is stopped using PMSF (phenylmethylsulfonyl fluoride) in a final concentration of $5 \times 10^{-4}$ M at an optical density of $OD_{600}$ of 1.0 and at an induction time of 1 hour, cooled in ice and spun down. The cell sediment is then washed in 1 ml of buffer (10 mM tris.HCl, pH 7.6; 40 mM NaCl), spun down and resuspended in 200 µl of buffer (20% sucrose; 20 mM tris.HCl, pH 8.0; 2 mM EDTA), incubated at room temperature for 10 minutes, spun down and immediately resuspended in 500 µl of double-distilled $H_2O$. After incubation in ice for 10 minutes, the shock-lysed bacteria are spun down and the supernatant is frozen. The proinsulin content of this supernatant is tested by a standard insulin RIA (Amersham).

The bacterial sediment is resuspended once more in 200 µl of lysozyme buffer (20% sucrose; 2 mg/ml lysozyme; 20 mM tris.HCl, pH 8.0; 2 mM EDTA), incubated in ice for 30 minutes, sonicated 3 x 10 seconds and then spun down. The supernatant resulting from this is tested for the content of proinsulin ("plasma fraction") in a radioimmunoassay.

Individual bacterial clones which contain the plasmid pWIP 1 were examined for their synthetic capacity and their ability to transport the proinsulin-presequence product. It was possible to demonstrate that all the bacterial clones, as expected, transported about 90% of the produced proinsulin into the periplasmic space. About 10% of the RIA activity of proinsulin was still found in the plasma fraction.

DNA sequence I

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Triplet No. | | | | | | | | 1 | | 2 | | 3 | |
| Amino acid No. | | | | | | | | Met | | Lys | | Gln | |
| Nucleotide No. | | | | | | 5 | | | | 10 | | | |
| Coding strand | | | 5' | AA | TTC | ATG | AAA | CAA | | | | | |
| non-cod. strand | | | 3' | | G | TAC | TTT | GTT | | | | | |

| 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|
| Ser | Thr | Ile | Ala | Leu | Ala | Leu | Leu | Pro | Leu |

| 15 | 20 | | 25 | | 30 | 35 | | 40 | |
|---|---|---|---|---|---|---|---|---|---|
| AGC | ACG | ATC | GCA | CTG | GCA | CTC | TTA | CCG | TTA |
| TCG | TGC | TAG | CGT | GAC | CGT | GAG | AAT | GGC | AAT |

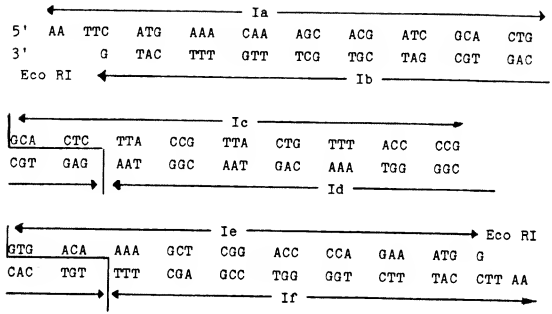
| 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 |
|---|---|---|---|---|---|---|---|---|---|
| Leu | Phe | Thr | Pro | Val | Thr | Lys | Ala | Arg | Thr |

| 45 | 50 | | 55 | | 60 | 65 | | 70 | |
|---|---|---|---|---|---|---|---|---|---|
| CTG | TTT | ACC | CCG | GTG | ACA | AAA | GCT | CGG | ACC |
| GAC | AAA | TGG | GGC | CAC | TGT | TTT | CGA | GCC | TGG |

| 24 | 25 | 26 |
|---|---|---|
| Pro | Glu | Met |

| 75 | 80 | | 84 | | |
|---|---|---|---|---|---|
| CCA | GAA | ATG | G | | 3' |
| GGT | CTT | TAC | CTT | AA | 5' |

DNA sequence II:



THE EMBODIMENTS OF THE INVENTION IN WHICH AN EXCLUSIVE PROPERTY OR PRIVILEGE IS CLAIMED ARE DEFINED AS FOLLOWS:

1. A synthetic signal sequence for the transport of proteins in expression systems which comprises DNA essentially corresponding to a natural signal sequence but having one or more cleavage sites for endonucleases which the natural DNA does not contain.

2. A signal sequence as claimed in claim 1, which contains internal cleavage sites upstream or downsteam or upstream and downstream of the hydrophobic region.

3. A signal sequence as claimed in claim 1, which essentially corresponds to the natural signal sequence of alkaline phosphatase of E. coli.

4. A signal sequence as claimed in claim 1, which contains at the 3' end up to about 40 of the amino-terminal codons of the adjacent structural gene following downstream.

5. A signal sequence as claimed in claim 2, which essentially corresponds to the natural signal sequence of alkaline phosphatase of E. coli.

6. A signal sequence as claimed in claim 2, which contains at the 3' end up to about 40 of the amino-terminal codons of the structural gene following downstream.

7. A signal sequence as claimed in claim 3, which contains at the 3' end up to about 40 of the amino-terminal codons of the structural gene following downstream.

8. DNA of the formula I:

```
                                          5          10
                                5'   AA  TTC  ATG  AAA  CAA
                                3'        G  TAC  TTT  GTT

15     20         25          30     35         40
 AGC  ACG  ATC   GCA   CTG    GCA  CTC   TTA   CCG  TTA
 TCG  TGC  TAG   CGT   GAC    CGT  GAG   AAT   GGC  AAT

45     50         55         60     65         70
 CTG   TTT  ACC  CCG  GTG  ACA  AAA  GCT  CGG  ACC
 GAC   AAA  TGG  GGC  CAC  TGT  TTT  CGA  GCC  TGG

75     80        84
 CCA  GAA  ATG   G        3'
 GGT  CTT  TAC   CTT  AA  5'
```

9. A process for the transport expression of eukaryotic, prokaryotic or viral proteins in prokaryotic and eukaryotic cells, which comprises coupling the gene for the protein which is to be transported onto a DNA sequence as claimed in claim 1, incorporating this fusion gene into a vector, and transforming therewith a host cell which transports the expressed protein out of the cytoplasm.

10. The process as claimed in claim 9, wherein the synthetic DNA signal sequence codes for a protein intrinsic to the host.

11. A hybrid vector comprising a DNA sequence as claimed in claim 1.

12. A hybrid vector as claimed in claim 11, which is a hybrid plasmid containing the DNA sequence I as claimed in claim 8, inserted in an Eco RI cleavage site.

13. A host cell containing a vector as claimed in claim 11.

14. A host cell containing a vector as claimed in claim 12.

15. A host cell as claimed in claim 13, which is of the species E. coli.

16. A host cell as claimed in claim 14, which is of the species E. coli.

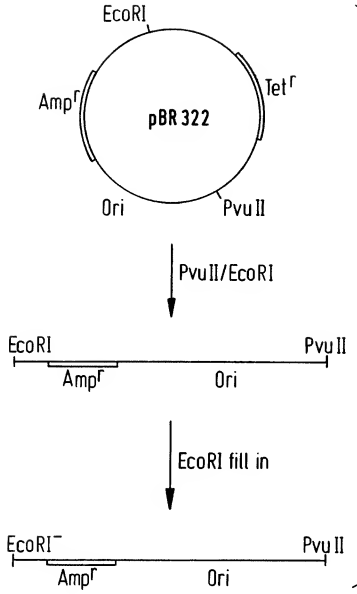



FIG.1

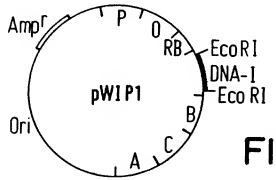
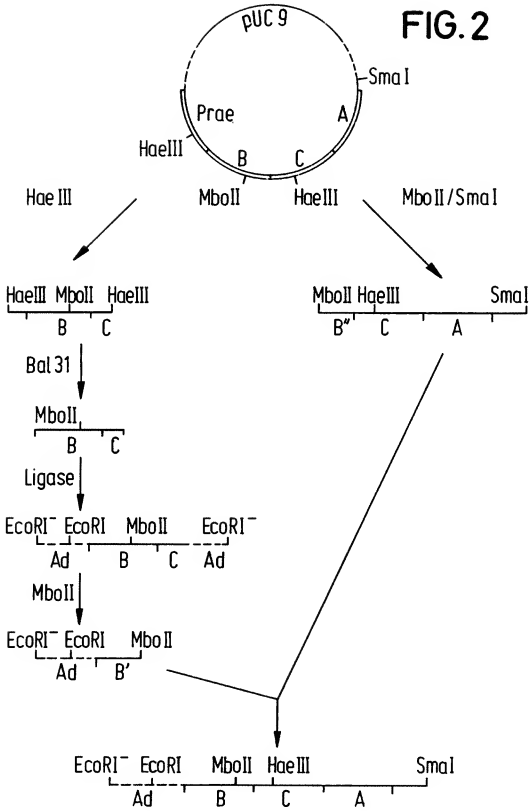



FIG.5

Ag: Regis, Bernstein & Plumb

FIG. 2




By: Rogers, Bereskin & Parr

1310230

FIG. 3

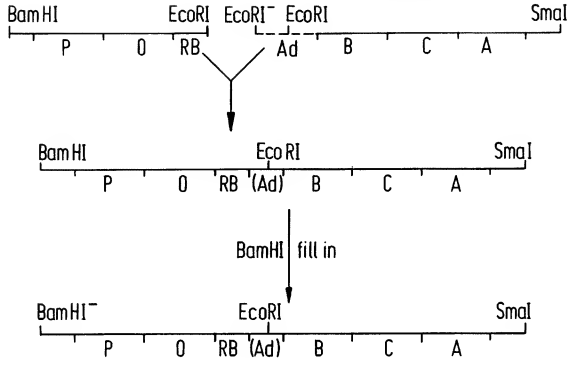
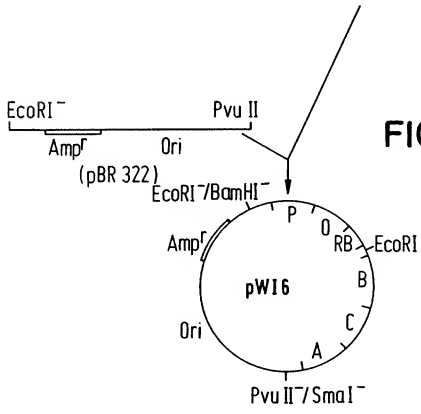



FIG. 4




Atty: Rogers, Bercowitz & Rose